

Exercise 7

Lecture 10: Database Analytics

Scalable Data Engineering

7.1 Are the following statements true or false? Give reasons for your answers.

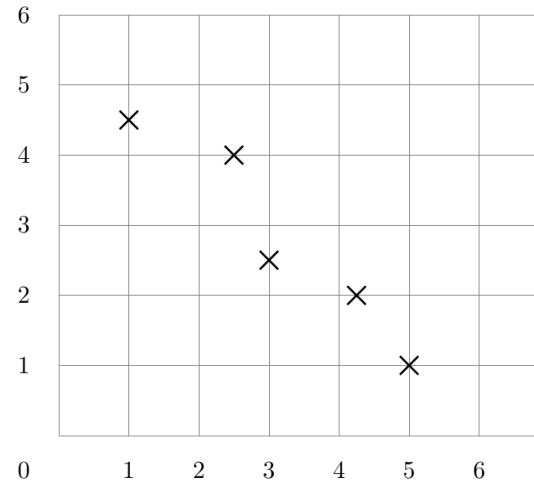
- a) Extending a data base system is pointless since data bases are capable of everything an analyst will ever need.
- b) There are no fundamental differences between functions and procedures.
- c) An aggregate function calls the step function one last time after all tuples are processed.
- d) Regression is the mapping from a feature vector to multiple categories.
- e) Regression does not suffer from overfitting.
- f) Overfitting is generally good because it guarantees a fitting model for all data and improves the model accuracy.
- g) The test set is a subset of the training set.

Ordinary Least Squares (OLS)

7.2 Use OLS to fit a linear function to the provided data.

- $y = mx + c$
- $\operatorname{argmin}_{\alpha} \sum_{i=1}^n (y_i - \vec{x}_i \alpha^T)^2$ with $\alpha = (m, c)$

$\vec{x}_i$	y_i
(1)	4.5
(2.5)	4
(3)	2.5
(4.25)	2
(5)	1



7.3 Use the `plpython3` extension of PostgreSQL to solve the real estate housing regression.

- Create table with example data.
- Access database data from python function.
- Solve OLS.
- Return slope and intercept.

slope	intercept
118.89	90229.56

size	price
1000	275000
2500	370000
800	175000
1900	225000
3000	500000